Cultural and Geolocation Aspects of Communication in Twitter

Cultural and Geolocation Aspects of Communication in Twitter

E. Daehnhardt[1], Y. Jing[2], N.K. Taylor[1]
School of Mathematical & Computer Sciences, Heriot-Watt University
School of Computing, London Metropolitan University
elena@daehnhardt.com, y.jing@londonmet.ac.uk, N.K.Taylor@hw.ac.uk

Abstract

Web applications exploit user information from social networks and online user activities to facilitate interaction and create an enhanced user experience. Due to privacy issues however, it might be difficult to extract user data from social network, in particular location data. For instance, information on user location depends on users' agreement to share own geographic data. Instead of directly collecting personal user information, we aim to infer user preferences by detecting behavior patterns from publicly available micro blogging content and users' followers' network. With statistical and machine-learning methods, we employ Twitter-specific features to predict country origin of users on Twitter with an accuracy of more than 90% for users from the most active countries. We further investigate users' interpersonal communication with their followers. Our findings reveal that belonging to a particular cultural group is playing an important role in increasing users responses to their friends. The knowledge on user cultural origins thus could provide a differentiated state-of-the-art user experience in microblogs, for instance, in friend recommendation scenario.

1 Introduction

Social networking sites such as Twitter microblogs allow users to communicate with their online friends and share information in real time. Some web and mobile applications require information on user location and origin to provide users with location-specific information, like recommending places of interest, social events or online friends in close proximity. A user profile containing user geographic locations, on which recommendations can be provided may impose also possible privacy threats. This is why a majority of Twitter users avoid sharing their accurate locations [1].

However, despite of user efforts to hide or obscure one whereabouts, there are methods to identify user origins based on content [2, 3, 4], social profile metadata [2, 5], and from other social networking sites, in which users' data is gathered [6]. Besides location-related cues, microblogging activity patterns can reveal users from different origins, which could be used for indirectly inferring origins [7].

Determining precise user locations is discussed in previous works on a city-level granularity [8]. Striving to preserve user privacy, we abstract from mining accurate location-specific information.

We limit ourselves to a country-level or a "cultural group" comprising a group of countries with further defined behavioral patterns. User similarities and differences found in the profile meta-data, contact networks, content and microblogging behavior can be employed as a proxy for finding user origins, whether they are country or culture-related. Our contributions to social networking research are the following:

- Investigation into the predictive value of Twitter user-related and social-network related features in experiments for predicting user origins.
- Analysis of user communication preferences in Twitter on the example of followers' responses to their friends.

In section 2, we outline related work in the scope of Twitter location detection, cultural aspects playing a role in adaptation and personalization. In section 3, we describe our research methodology and experimental setup. Next, we outline and discuss the results in user origin and response prediction experiments. We also discuss benefits and limitations of the proposed approach and possible areas of improvement. We conclude with research findings, how our findings could be used in the development of adaptive social web applications.

2 Related Work

One of the main requirements for adaptation is user location, which can be used to mine location-specific content [9] and user interests [10]. Explicit user location, however, is often missed or not accurate, and only a very small part of openly-available microposts are geographically tagged [1]. This is why location detection out of social web is a pertinent research topic.

A method to detect locations from content is the usage of a gazetteer or toponym vocabulary comprised of location-specific terms. However, the application of gazetteer for location information disambiguation in microblogs is challenging for similarly named locations and the inherent difficulties of mining information out of the microblogs. Misspelling and usage of abbreviations is common due to the short message limitations [11]. To improve performance of location detection using GeoNames gazetteer, [11] employ Support Vector Machines classifier using Twitter features extracted from tweet content and meta-data.

As assessed by human annotators, geo-coding services Yahoo and Google applied to profile locations on Twitter are not really trustful for a large proportion of tweets [12]. This is not surprising since about 30% of users do not provide an accurate location [1]. As [2] pointed out, user location is influenced by temporal dynamics and requires a model update when used as a feature in a location detecting classifier. Named-entity recognition with Stanford NER and Open NLP tools is investigated in [13], showing a considerable performance when training on Twitter-specific content, which requires human involvement for annotating data. Location disambiguation of Tweets with Stanford NER, gazetteer, heuristic rules was performed with precision and recall of around 80%, and is comparable with human annotations [3]. Authors also suggest that representation of the tweets' content with help of the ontologies [3] might be useful in the toponym disambiguation task.

The detection of the home country from Twitter content was also investigated in [1] with machine learning technique whilst analyzing tweets' content, disregarding other Twitter features and the extended contacts' network of the users. Using Naive Bayes classification model working with term frequencies, user country locations were inferred in about 73% of cases [1]. Geographical topic models created based on terms extracted out of content correlate with specific geographic locations, but require an adjusted probability estimation with help of smoothing technique in order to deal with term sparsity [14, 10, 4]. Location-specific terms selection with Kernel Density Estimation is investigated in [15], demonstrating robustness of the approach when only a small set of terms is employed.

[2] exploit location, user name, description and time zone fields for creation of a location-detecting classifier, finding country location with an accuracy of 92% for the best feature set analyzed. [8] created an ensemble classifier for detecting users' home location based on words and hashtags extracted from tweets, tweets' frequency dynamics and gazetteer dictionary of geographic place names. Their classification algorithm enabled hierarchical location detection of time zone, state and city name with recall figures of 0.78, 0.66 and 0.58 respectively. For improving location detection outcomes for Flickr images, [5] use statistical inference, gazetteer and other features extracted from the Flickr users' meta-data and content.

In location disambiguation based on user-generated content, adding content of contact users helped to improve prediction performance [16]. Locations of users sharing web links help to predict the link origins [17]. Locations of users can be predicted based on their contact networks, also considering different social platforms [6].

Observing Twitter connections within national and international geographical boundaries, [18] discuss the nations as defining cultural representations of communities, in which people communicate differently due to their different interests and geo-political status of their countries. There are patterns of follower-friend relationships, in which US users are usually followed by users from other countries, while Japanese are usually followed mostly by Japanese [18]. Interestingly, the majority of connections are within national borders [19, 18]. In accord with [18], 39% of connections are within 100 km distance. English-speaking users follow also English-speaking users in the majority of over 90%, while other language users build connections with users of the same language, in 60% of cases [18].

There are several research domains investigating adaptation and personalization outcomes considering cultural backgrounds of users. In e-learning, [20] found that the cultural context of a learner might impose requirements on technology usage, selection of media and the style of interaction between students and instructors. The learning material presentation methods impact the performance of students from different cultural groups [21]. To improve the experience of e-learners in e-learning environments, it is paramount to distinguish between different cultural preferences [20].

In a recommendation system context, [22] analy-

Table 1: Research Question and Hypothesis

RQ1	How can Twitter microblogs be exploited to infer user origin and locality?
H1.1	The information on country location derived from user tweets' meta-data and respective meta-data of the followers is not sufficient for providing cues on user origins for the majority of users (75% from our dataset).
RQ2	Which Twitter-specific features could be used for inferring the country location of users?
H2.1	A users contact network can assist in improving country prediction.
H2.2	User microblogging patterns can assist in further improving locality prediction.
RQ3	Which friend features are important for predicting a user's follower responses?
H3.1	User and follower's country locations' and language match are amongst the most important prediction parameters for user responses.
H3.2	User's influence is significant in predicting her follower responses.

ses recommendation prospects for urban planning in accord with user cultural similarities on Foursquare and user preferences. Cultural behavior differences in user activities on Twitter were investigated in [23], suggesting to exploit such differences for building responsive communication applications. Other possible applications include friendship recommendations based on suggestions by the network [24] and a community detection approach based on user interactions on Twitter [25]. Microblogging response prediction with focus on English-speaking users was studied in [26], which analyzed the importance of specific terms occurred in tweets, previous response ratios and also number of user and follower links in the social network. The importance of social relationship between Twitter users on their responses prediction revealed in [27].

Overall, the above mentioned research points out cultural behavior differences of users online. It remains however unclear whether knowledge on the user cultural background would help in improving recommendation performance, and whether such recommendations would outperform country-specific recommendations. A further in-depth investigation into follower and friend relationships on Twitter could shed a light into location-specific cultural aspects playing a role in online communication. For predicting user responses, we further evaluate the inclusion of the geographic and culture-specific aspects of the users involved. For this, we examined the user and follower-related features enabling to identify user country and culture-level origins, which we further exploited in a user response (reply or retweet cases) prediction experiment.

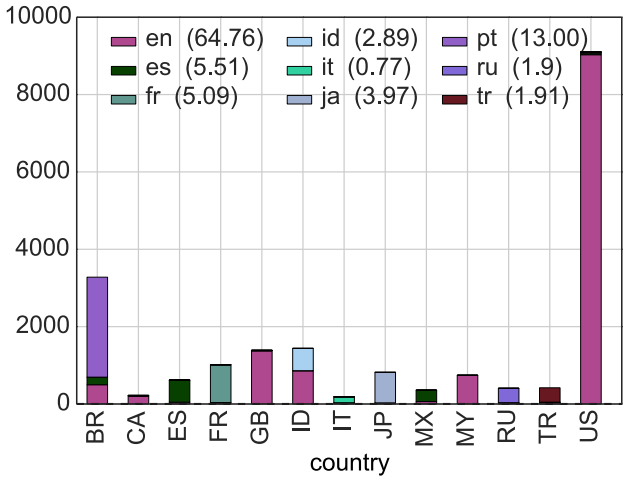
3 Methodology

The main aim of the study was to analyze user communication in Twitter, in order to get insights into friend-follower relationships for further reply predictions. It seems reasonable to assume that user geographic locations, cultural origin and language might play a role in the follower interests reflected in reply or retweet messages.

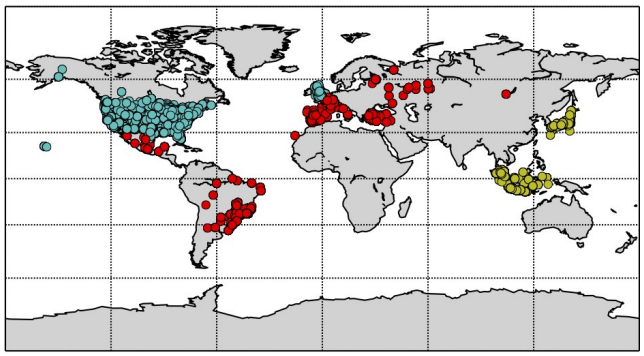
We also analyzed the relative importance of other Twitter-specific features. For instance, number of user friends and followers, which ratio is often explained as an "Influence" on other users might help us to detect interesting users to follow. Also, we assess user-related and follower-related features for predicting a user whereabouts and the interest of the followers. We present our research questions and hypothesis in Table 1. Our research is based on the following assumptions and definitions:

- cultural group is the group of individuals comprising one or several nations exhibiting similar behavior and communication traits as previously explored in the sociological studies such as Hofstede, which application in information system research is critically assessed in [28]. We are aware that not all individuals are expected to exhibit behavior strictly within an associated cultural or country group. However, we employ the cultural group notion to merely stereotype online users.

- Based on previous research [7], we employed the Lewis model of cultures describing how persons belonging to different cultural backgrounds differ in their interpersonal behavior. For instance, Asian cultures, such as Japan or Vietnam, are defined by Lewis in their culture dimension as Reactive (RE), since they are generally considered to be courteous, accommodating and good listeners. In addition, Lewis defines a Multi-Active (MA) and Linear-Active (LA) cultural dimension. While persons described as MA focus on interpersonal communication and are generally considered as emotional personalities, LA



(a) Languages (percentages) and Countries (ISO codes), the Vertical Axis Shows the Number of Users



(b) User Locations and Dimensions (blue: LA, yellow: RE, red: MA)

Figure 1: Top Languages and Geographic Locations

persons focus on working with facts and planning activities [29].

- Therefore, each of the cultural groups/nations can be described with the help of "cultural dimensions". In this paper, we use cultural group and dimension terms interchangeably. To create three main cultural groups, including MA, LA and RE, we combine users from several countries. With human annotation, dimension codes where assigned to the countries which were analyzed, as seen in Figure 2.

Country (Code)	Dimension	Country (Code)	Dimension
Brazil (BR)	MA	Mexico (MX)	MA
France (FR)	MA	Russian Federation (RU)	MA
Spain (ES)	MA	Italy (IT)	MA
Turkey (TR)	MA	Indonesia (ID)	RE
The United States (US)	LA	Japan (JP)	RE
Great Britain (GB)	LA	Malasya (MY)	RE
Canada (CA)	LA		

Figure 2: Countries and Assigned Dimension Codes

3.1 Experimental Setup

Selected Users. Using Twitter Streaming API, we collected a sample of of 4250109 tweets published by 3198307 users in the period from 17th to 18th of March 2014. Only about 2% out of these tweets were provided with geographic locations. From this "Sample" dataset, we randomly selected 20.000 users with published tweets containing geographic locations as seen in Figure 1. The geographic information available in the tweets' meta-data helped to reduce preprocessing time. In order to accurately determine a user's origin, it was paramount to decrease as much of possible the number of users travelling or residing in countries other than their country of origin. For this, we introduced a parameter α, which equals one when user language defined in the user profile matches with the top first native language related to the user's country, and zero otherwise. This requires however a large number of users when training our classification models for improving accuracy when identifying users and their respective origin.

User Profiling Features. For the selected users, we followed their tweets, replies and retweets of their followers in the period from 18th to 25th of March 2014. The "Follow" dataset comprised these tweets, in which around 31% of 2853719 tweets are geographically tagged and originated from the initially selected users and 131174 of their followers.

For each of the pre-selected geo-tagged users, we created feature vectors representing a summary of users' activity summaries, which were stored into the dataset "Profiles". The "Profiles" dataset consists of the 13289 pre-selected users defined as coming from the top most active countries in our sample dataset, namely the USA (US), Brazil (BR), Indonesia (ID), Great Britain (GB), Turkey (TR), Japan (JP), France (FR), Spain (ES), Malaysia (MY), Mex-

ico (MX), Russia (RU), Canada (CA), Italy (IT) and where content from followers was available. We exploited the following features for detecting user country origin:

- LANGUAGE User **Language** from user's Twitter profile.
- BEHAVIOR features set included features related to user activities on Twitter.
    1. **Tagging**: number of Hashtags divided by the sum of Hashtags and Uniform resource locators (URLs) occurring in user-generated content; denotes users' preferences towards tagging and sharing content activities.
    2. **Languages**: number of different languages detected from user content[1], normalized by division of the mean values of languages employed by all users.
    3. **Weekends**: number of tweets published on weekends divided by number of tweets published by the user; denotes frequency of posting during weekends.
    4. **Response**: number of user replies divided by the total number of user replies and retweets.
    5. **Mentions**: defines user preferences for sharing information on other users as compared to sharing Hashtags and URLs; calculated from the total number of user mentions divided by the sum of URLs and Hashtags. This feature reflects users' focus on people or organizational activities as described in Lewis [29].
    6. **Mobility**: denotes the number of different country occurrences in the user tweets' meta-data divided by the mean value of the number of country occurrences in the tweets of all users.
    7. **Timezones**: number of different time zones in the tweets' meta-data of a user.
    8. **Influence**: number of Followers divided by the total number of Followers and Friends.
- META. The text string **Meta** is created by joining strings of the language code defined in the user profile, most used Time zone and Location found in the tweets' meta-data.
- CONTENT: for each user one tweet's text content is extracted.
- LOCATION: location field found in the user meta-data.
- FOLLOWERS features set consists of features extracted from the user follower networks.
    1. **FCountry**: the country mentioned most in followers' meta-data.
    2. **FCountries**: the number of different countries referred to in the followers' meta-data.
    3. **FLanguage**: the language most mentioned in the followers' tweets' meta-data.
    4. **FLanguages**: number of different languages found in the followers' tweets' meta-data.
    5. **FTimezone**: time zone most referred to in the followers' meta-data.
    6. **FTimezones**: number of different time zones referred to in the followers' meta-data.
    7. **FInfluence**: number of Followers divided by the total number of Followers and Friends for each of the user Followers, further taking the mean value for the user we follow.

For determining a user country location, we employed Twitter meta-data and Twitter specific elements. We did not apply named-entity recognition algorithms, since they are resource-demanding and some named entities may be quite ambiguous in distinguishing from other words [30, 31]. The initial feature choice was guided by the literature sources and our experimental system design. In addition to content and user-related features, we also included follower-related behavior, to examine a feature's importance in relation to country detection performance. In further experiments, we also examined the user responses in friend-follower relationships.

Country Detecting Classifier. Each of our initially selected users had an associated country code (string with ISO 3166-1 alpha-2 country code) found in the user meta-data of tweets. We used these users for training our classification model based on the features described in this section.

To each of LANGUAGE codes we assigned a numerical value, while the BEHAVIOR feature set included only numerical values. In the FOLLOWERS features' set, we coded FCountry, FLanguage, FTimezone as numerical values, while the rest were computed as integers (FCountries, FLanguages, FTimezones) or real values (Influence). When dealing

[1]Python library "langid": https://github.com/saffsd/langid.py

with numerical values, for building our classification models, we exploited the Decision Trees Classifier, which also allowed to consider the importance of a feature.

When dealing with language features extracted from the user tweets (CONTENT, LOCATION and META), we created pipelines performing the following steps:

- Convert text data such as tweets' content or location field to a matrix of token counts;
- Convert the count matrix into its normalized Term Frequency - Inverse Document Frequency statistics;
- Employ the Multinomial Naive Bayes classifier for predicting user countries.

For evaluating our country classification results, we performed a three-fold Cross-Validation (CV), and employed Accuracy and F1-measure [32]. For creating our training and test samples, we split the "Profiles" dataset into fractions of 75% and 25%.

Communication Patterns and User Responses. Based on the "Follow" dataset, we created the "Communication" dataset representing 107960 of user tweets, replies and re-tweets, from which about were 88% geographically tagged and published by the pre-selected users and their followers. We then analyzed user interactions among different user groups. Next, we created and evaluated a response predicting classifier based on the ratings computed using the "Communication" dataset as follows:

- For each of the pre-selected users, we computed the number of their followers' retweets and replies, which we exploited for creating a rating score ranging from 0 (no replies and retweets for a particular user and follower combination) to 1 (the maximal number of retweets and replies for user and follower communications). We rated each pre-selected user in a way to assess the interest for a particular follower.
- We predicted user responses towards their friends by training our decision tree and logistic regression models based on the aforementioned features and a set of features denoting users' matches in the language usage and location.

4 Results

4.1 Explicit Locations in Meta-data

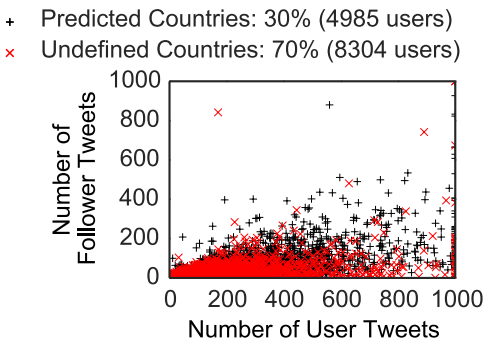


Figure 3: Countries Detection Test

We were interested in evaluating how useful country-related information extracted from user tweets' and followers' tweets is. The aim was to find out the required amount of Twitter-content to infer country-locations of users in our dataset and assess the possibility of detecting user origins based on the meta-data of tweets of the user and followers.

Our first hypothesis states that meta-data of user and followers tweets are not sufficient for explicitly inferring country locations for a majority of selected users. The average number of tweets per user is about 119 tweets (Standard deviation 156 tweets), while the number of friends and followers is about 590 (St.dev. 2731 friends) and 894 (St.dev. 7598 followers) respectively per user in average. Even though the fraction of geo-tagged tweets from our "Follow" dataset exceeds the randomly selected tweets from the "Sample" dataset in more than 15 times, reaching around 31%, the amount of tweets collected during one week was not sufficient to infer country-specific information for about 70% of the users, as shown in Figure 3. In spite that geographically-tagged tweets occur only in about 2% of cases in our "Sample" dataset, the location field enabling users to specify their locations arbitrary in text was filled in 54.6% of cases. However, here we might disregard the location field (further examined below in Table 2) and usage of geo-coding services due to the ambiguity and often use of the location field by Twitter users for personal or humoristic comments, as stated in [1] in more than 30% of cases. Therefore, we assert in our H1.1 hypothesis, that for the majority of cases the country location meta-data is not sufficient for inferring user countries based on followers' and user tweets meta-data alone.

4.2 User Country Prediction

Next, we analyzed how the different combination of features can be useful for distinguishing between dif-

Table 2: Performance of the Country-detecting Classifier (in test/train split: 4650 training and 1550 testing instances, CV Acc.: 3-times cross-validation accuracy for 6200 instances in training and test datasets in total)

	CV Acc.		Accuracy		Precision		Recall		F1	
Feature Set	α=1	$\forall\alpha$	α=1	$\forall\alpha$	α=1	$\forall\alpha$	α=1	$\forall\alpha$	α=1	$\forall\alpha$
User-related Data										
LANGUAGE	**0.88**	0.76	0.88	0.76	0.78	0.70	0.88	0.76	0.82	0.71
LOCATION	0.62	0.58	0.64	0.59	0.63	0.66	0.64	0.59	0.53	0.51
META	**0.91**	0.85	0.91	0.85	**0.90**	0.86	0.91	0.85	**0.90**	0.83
BEHAVIOR+LANGUAGE	0.80	0.66	0.81	0.66	0.81	0.67	0.81	0.66	0.81	0.67
CONTENT	0.63	0.58	0.65	0.58	0.55	0.45	0.65	0.58	0.54	0.47
BEHAVIOR	0.44	0.38	0.46	0.38	0.48	0.39	0.46	0.38	0.47	0.39
Follower-related Data										
FOLLOWERS	0.88	0.84	0.88	0.84	0.88	0.84	0.88	0.84	0.88	0.84
Mixed Data										
LANGUAGE+FOLLOWERS	**0.94**	**0.87**	0.94	**0.87**	0.94	**0.87**	0.94	**0.87**	0.94	0.87
BEHAVIOR+FOLLOWERS	0.87	0.82	0.88	0.83	0.88	0.83	0.88	0.83	0.88	0.83
BEHAVIOR+FOLLOWERS+LANGUAGE	0.92	0.87	0.91	0.87	0.91	0.87	0.91	0.87	0.91	0.87

ferent countries of user origins.

User Data. Our aim was to achieve an accuracy of above 46% of country prediction when considering the majority class classifier's threshold, since about 46% of users in the "Profiles" dataset originated from the USA, with English defined in their user profiles. Moreover, in our dataset some of the countries such as Indonesia (ID) and Malaysia (MY) have a large fraction of English language users, even though English is not the native language. This is why some of the users are misclassified if only the language defined in their user profiles is considered. However, our findings revealed that the LANGUAGE classification strategy outperformed all other User-related strategies, for the exception of META, in respect to CV accuracy, as seen in Table 2.

The META feature is represented by a text string comprising language defined in user profile, majority time zone and location. With the META feature, we achieved 91% CV accuracy (three-fold CV showed the best CV accuracy performance in all our tests), which outperformed the performance of the LANGUAGE's strategy. Since the location-specific information is not always accurate [1], we further analyzed other feature mixes. The CONTENT strategy slightly outperformed LOCATION-based strategy in $alfa = 1$ cases for Accuracy, Recall and F1 measures while enabling to achieve 63% CV accuracy when based on only one tweet's content per user. The BEHAVIOR strategy performed poorly compared with other classification strategies, however, we noticed a slight precision improvement in $alfa = 1$ cases with lower recall and accuracy values when using BEHAVIOR together with the LANGUAGE feature. It seems that the BEHAVIOR feature does not allow us to improve user classification into country groups when using only user-related data. Considering a relatively small number of users in our "Profiles" dataset, we were not surprised to achieve only 44% of CV accuracy and 46% test accuracy when using the BEHAVIOR feature set, which might require more data and features to yield similar results as the strategies analyzed above.

Followers Data. In cases when user metadata was not available or deemed to be inaccurate, the FOLLOWERS strategy could compete with some User-related feature sets. We achieved more than 5% improvement in Precision and F1-measure compared to all User-related feature combinations, with exception of META. Interestingly, in a majority of test cases, we observed better performance when considering $\alpha = 1$ cases. We might reasonably assume, that the combination of user language and user country of origin is important for selecting training instances. Overall, the FOLLOWERS feature set provided a viable alternative for detecting user countries in our experiments when accurate META and LANGUAGE data was absent.

User and Followers. When combining user LANGUAGE with information extracted from the followers' network, we achieved the best performance in all measures, in all our experiments for $alfa = 1$ cases. The cross-validation accuracy of LANGUAGE+FOLLOWERS combination was around 94%. Therefore, we might accept our hypothesis H2.1, that the user contacts' network improves user country predictions. Despite of our expectations, combining BEHAVIOR with FOLLOWERS features did not improve classification performance compared

Table 3: Relative Features Importance (User Country Prediction)

BEHAVIOR+LANGUAGE+FOLLOWERS							
Feature	Importance (%)	Feature	Importance (%)	Feature	Importance (%)	Feature	Importance (%)
Language	100	FCountry	16.16	FTimezone	15.36	FInfluence	2.34
Weekends	2.34	Influence	2.13	Mentions	1.55	Tagging	0.98
Response	0.73	FTimezones	0.51	Timezones	0.36	Languages	0.33
FLanguage	0.32	FLanguages	0.26	FCountries	0.15	Mobility	0.02

to using only FOLLOWERS. We could not accept the H2.2, that the BEHAVIOR patterns could help in improving user country predictions in our experimental settings. Overall, when using Follower-related features and LANGUAGE, we were able to outperform the accuracy when using only LANGUAGE/META features and also the Calgari algorithm using Naive Bayes classification model, as described in [1].

Combining all non-content features in the "BEHAVIOR+LANGUAGE+FOLLOWERS" classification strategy enabled us to assess the relative importance of features for country detection using Decision Tree classifier. Our experimental results showed that user language, followers' majority country and time zone, followers' average influence and posting day were the most important features for detecting user country of origin, while the user mobility was the least important feature to consider, as seen in Table 3.

4.3 Predicting Follower Responses

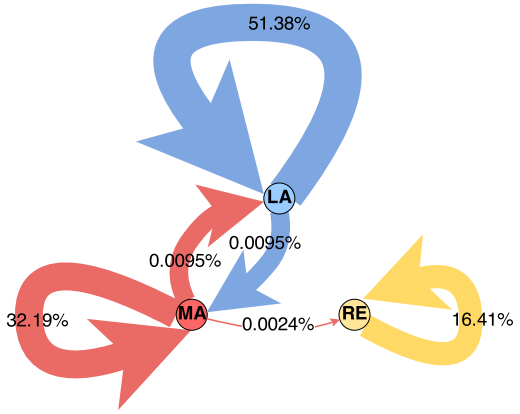


Figure 4: Communication between Cultural Dimensions for geographically-tagged users

Our analysis of communication between users and their followers demonstrated that a substantial proportion of users follow other users from the same cultural groups as seen in Figure 4. This might indicate, that our users are more interested to follow others in their specific area of interest, in particular countries.

Next, we investigated whether the dimension or country location of user and follower match, as well as other user-related features having an influence on communications between Twitter friends. For this, we created a classification model based on the decision tree and logistic regression techniques while using the aforementioned 16 features. We also added binary variables such as "LangMatch","CountryMatch","DimMatch" (True when user and followers' profile languages, Countries and Dimensions match), "FLangMatch" (True when user followers' and followers followers' profile languages match), "FCMatch" (True when user followers' and followers followers' Country match), "FTimezMatch" (True when user followers' and followers followers' time zones match).

When a followers' location was unknown, we employed the previously constructed classification models "META" and "LANGUAGE". The parameters "CountryMatch" and "DimMatch" were set to True when friends' parameter matched with the parameter of the follower based on one of the three values: value from the profile information, parameter's value derived using "META" or "LANGUAGE" country detection models, or employing "LANGUAGE" model for detecting a dimension value directly.

To find out what is important for predicting user responses to the friends of a user, we performed the following steps:

- For each of the initially followed users, we calculated a sum of follower responses (retweets and replies) for each of their followers, which was further normalized by dividing it with the maximum sum of responses for a follower. This way we ranked our initial Twitter users with numbers from 0 (no response) to 1 (maximum number of response) to account for the "interestingness" of each of a user's followers.

- Since a large fraction of our followers had very few replies, our ratings included more than 98%

Table 4: Relative Features Importance (RFI) in Followers' Response Prediction Test using Decision Trees, Logistic Regression Analysis Results (Predicted Logit of Interest) and Logit Marginal Effects, statistically significant (with $p < 0.05$) are shown in bold font

Parameter	D. Tree	Logistic Regression Results, pseudo $R^2 \approx 0.23$, sensitivity ≈ 62, specificity ≈ 72							Marginal Effects					
	RFI	Odds Ratio	z	$P > \|z\|$	β	Std. Err.	95% Conf.Int.		dy/dx	Std. Err.	z	$P > \|z\|$	95% Conf.Int.	
Intercept		172.7	2.61	0.01	5.15	1.97	1.29	9.02						
CountryMatch*	2.22	0.46	-1.48	0.14	-0.78	0.53	-1.82	0.25	-0.14	0.09	-1.49	0.14	-0.33	0.04
DimMatch*	3.91	4.45	2.31	0.02	1.49	0.64	0.23	2.76	0.23	0.11	2.35	0.02	0.04	0.49
FCMatch*	9.63	0.68	-1.96	0.05	-0.38	0.19	-0.76	-0.00	-0.07	0.03	-1.98	0.05	-0.14	-0.00
FLangMatch*	100	0.09	-10.03	0.00	-2.41	0.24	-2.88	-1.94	-0.43	0.03	-14.23	0.00	-0.49	-0.37
FTimezMatch*	7.20	1.20	0.86	0.39	0.18	0.21	-0.23	0.59	0.03	0.04	0.86	0.39	-0.04	0.11
LangMatch*	6.50	0.79	-0.67	0.50	-0.24	0.35	-0.93	0.46	-0.04	0.06	-0.67	0.50	-0.17	0.08
FCountries	24.79	1.16	1.44	0.15	0.15	0.11	-0.05	0.36	0.03	0.02	1.45	0.15	-0.01	0.06
FInfluence	74.49	0.26	-2.15	0.03	-1.36	0.63	-2.59	-0.12	-0.23	0.11	-2.17	0.03	-0.46	-0.02
FLanguages	56.84	0.85	-3.05	0.00	-0.16	0.05	-0.26	-0.06	-0.03	0.01	-3.12	0.00	-0.05	-0.01
FTimezones	25.78	0.98	-0.76	0.44	-0.02	0.03	-0.07	0.03	-0.00	0.01	-0.77	0.00	-0.01	0.01
Influence	70.61	0.94	-0.08	0.93	-0.06	0.73	-1.49	1.37	-0.01	0.13	-0.08	0.93	-0.27	0.25
Languages	23.37	0.78	-2.88	0.00	-0.24	0.08	-0.41	-0.08	-0.04	0.01	-2.94	0.00	-0.07	-0.01
Mentions	2.08	0.64	-0.62	0.53	-0.45	0.72	-1.86	0.96	-0.08	0.13	-0.62	0.53	-0.34	0.17
Mobility	2.96	0.14	-1.78	0.07	-1.97	1.10	-4.13	0.19	-0.35	0.20	-1.8	0.07	-0.74	0.03
Response	28.60	1.08	0.05	0.96	0.08	1.59	-3.04	3.20	0.01	0.29	0.05	0.96	-0.55	0.58
Tagging	27.93	1.11	0.31	0.75	0.10	0.32	-0.53	0.73	0.2	0.06	0.31	0.75	-0.09	0.13
Timezones	10.82	0.69	-1.65	0.10	-0.37	0.22	-0.81	0.07	-0.07	0.04	-1.66	0.09	-0.14	0.01
Weekends	77.61	2.04	1.37	0.17	0.71	0.52	-0.30	1.73	0.13	0.09	1.38	0.17	-0.05	0.31

with the highest rank of "interestingness". Out of 106775 ranks, only 343 ranks were of 0 value (no interest). This is why for creating our model based on the 343 ranks of 0 value (no interest), and 343 ranks of 1 value (highest interest), we disregarded other ranks.

- Next, having the Rank values as dependent variables indicating user response to a user's friend, user-related and follower-related features, we created our classification models.

Decision Trees. Having 686 instances in our "Interests" dataset, we splited it into training (75%) and testing (25%) sets for evaluating our classification model based on the decision trees technique. It achieved 65% of accuracy, 69% precision, 63% of recall and 65% of F1-measure. Even though the classification accuracy using "leave-one-out" cross-validation technique, was outperforming a random classifier in only about 15%, it enabled us to calculate variables' importance presented in Table 4 (Features Importance column) based on the training set of 685 instances.

Surprisingly, features relative importance statistics based on decision trees presented CountryMatch, DimMatch and LangMatch within the five least important features set. This is why we could not accept our Hypothesis H3.1, stating that user and followers country locations and languages match are amongst the most important prediction parameters for the user responses. The most important features were FLangMatch, Weekend, FInfluence, Influence, and FLanguages. It seems that Country and Dimension match were not as important for the user response predictions, when using decision trees classification model. We explain this by the possible biases in our dataset towards the most active users, including also broadcasting agencies. To further assess variables likelihood and effect on users' interest towards their friends, we perform logistic regression and compute their marginal effects.

Logistic Regression. We performed logistic regression analysis with the Statsmodel Package[2]. Table 4 presents logistic regression results considering binary dependent variable of user interest in user's friend coded as 0 (no responses) and 1 (most of responses). Some of the explanatory variables (marked

[2]http://statsmodels.sourceforge.net/

with *) where categorical and coded as True or False when the compared values were matched or not. The overall model was statistically significant with log likelihood ratio p-value less than 0.001. The pseudo R here cannot be interpreted as a measure of variance such as in a least squares regression.

The logistic regression showed the significance of the FLangMatch and FInfluence included in the top most important features derived with decision trees. Interestingly, DimMatch and Languages were also statistically significant. Marginal effects statistics presented in the last three columns in Table 4 showed that DimMatch is associated (statistically significant) with 23% higher probability of user response. FLangMatch, FInfluence and FLanguages are related to statistically significant lower probability of replying in 43%, 23% and 3% respectively. However, Influence was statistically insignificant in our logistic regression model. This is why we could not strictly accept our Hypothesis H3.2.

Hypothesis Revisited and Discussion. One of the objectives of this work was to investigate the possibility of exploiting microblogs' to detect a home country of a user. For this, we considered several countries on Twitter, whose users were deemed most active in our sample set. Firstly, we observed that country-name metadata of Twitter users was matched with the country-name meta-data of their followers in only 30% of our users. Therefore, we agreed with our H1.1 hypothesis stating that in most cases a country location taken from a user and followers' meta-data is not sufficient for providing cues on user origins in our dataset. Secondly, we found out that the information publicly available in user meta-data and followers' network enabled us to predict user country locations with a considerable accuracy of 90% or more for the best feature selection strategies we analyzed (RQ1). The most successful feature combinations included both elements, Followers-related data FOLLOWERS, and User-related data LANGUAGE (RQ2).

However, usage of BEHAVIOR together with LANGUAGE and FOLLOWERS feature sets did not provide any improvement. Therefore, we cannot strictly accept H2.2 without considering other features taking part in the classification strategy. Nevertheless, a solution to address the need for user profiling for improved user experience online, we might suggest exploiting user behavioral patterns or other well-performing features set combinations instead of directly asking user locations. This way, we could satisfy user preferences towards sharing content, times and ways of communicating with other users, whilst respecting privacy.

Overall, our results reveal a global orientation of our Twitter users in our dataset. Country Match and Language Match were not highly ranked (relative importance in decision trees), neither statistically significant features (logistic regression results). This is why we could not accept hypothesis H3.1. Dimension Match was more important than Country Match, and also statistically significant, leading to improved probability of user replies. Interestingly, user Influence was ranked in the top of feature importance in our decision tree results, while logistic regression showed no statistical significance, and FInfluence was even more important and significant. Thus, users with more influential followers might get less replies. When users have followers, with matching majority language, their reply probability decreases by 43%, which still supported the finding on the global nature of communication on Twitter. However, DimMatch could not be underestimated and requires a further investigation in further friend or related content recommendation experiments.

Privacy. Even though the country and user cultural dimension detection experiments open possibilities for adaptation, also concerns in regard of privacy issues could be raised, in particular for users with open profiles in Social Networks. Microblogging meta-data, followers' network and user-generated content enabled us to predict user country locations with considerable accuracy. Avoiding sharing location information in Twitter meta-data might help in preserving user whereabouts only to a certain extent. Even language mined from user content or defined in the user profile provides an insight into human countries of origin. Therefore, for a privacy-concerned user, we recommend to withdraw from microblogging or closing open profiles.

Sampling Biases. It is important to mention that our data collection method is prone to sampling biases. Using Twitter sample, we might be biased towards the most active users such as event or news broadcasters, which requires further analysis.

On Sociological Models Usage. The Internet brings users from diverse cultures together, however, understanding their needs and requirements in order to realize quality features for web applications is challenging. Applying sociological models to assess web site quality as perceived by users might not be trivial due to globalization, since the new e-culture of individuals often does not comply with rules described in models referring to differences in cultural personality [33]. Therefore, we might require new approaches to study user behavior online while respecting privacy.

Limitations. Our sample set contains users from the most active countries, providing geographical lo-

cations in meta-data. The user-generated content was collected for a period of one week. It is reasonable to assume that real-life activities might affect user behaviors. This is why we plan to explore user microblogging activities while assessing different models and feature combinations for predicting user origins and communication patterns in order to evaluate our approach in a larger time frame and extended locations set. For a thorough evaluation of our approach in predicting user responses to their friends, we might also involve human assessors working with our online version of the prototype.

5 Conclusions

In this paper, we analyzed microblogging activities for persons from the top 13 most active countries on Twitter. We investigated different feature sets extracted from the microblogging content and metadata of publicly available tweets. Our findings reveal that combining user-related microblogging features and features extracted from a followers' network enables user country prediction with an accuracy of more than 90%. We reflected on the results, also in view of human privacy issues, and provided recommendations for users concerned about sharing their data in open microblogs. Considering sociological studies and previous works on behavioral differences online, we proposed an approach for mining individual culture-specific microblogging preferences which we abstracted from country information, often revealed in user meta-data and content. This provides insights on the adaptation for web applications' and personalization options in order to preserve human privacy, while improving user satisfaction online by providing application features/content which are of interest for the cultural origins of the user. Finally, we investigated user interactions, and found out that users from the same cultural groups tend to communicate more with each other. In our next paper, we will analyze user communication amongst cultural groups in the long term, aiming to uncover recommendation approaches for further improving user experience on the Web. The other Twitter-specific features we might explore include tweeting frequency, topics found in the tweets' content, and an in depth tagging behavior analysis.

References

[1] B. Hecht, L. Hong, B. Suh, and E.H. Chi, "Tweets from Justin Bieber's Heart: the Dynamics of the Location Field in User Profiles", in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, 237–246.

[2] B. Han, P. Cook, and T. Baldwin, "Text-based Twitter User Geolocation Prediction.", *J. Artif. Intell. Res.(JAIR)*, 2014, vol. 49, 451–500.

[3] L. Ghahremanlou, W. Sherchan, and J.A. Thom, "Geotagging Twitter Messages in Crisis Management", *The Computer Journal*, 2014, bxu034.

[4] Z. Cheng, J. Caverlee, and K. Lee, "You are Where you Tweet: a Content-based Approach to Geo-locating Twitter Users", in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, 2010, 759–768.

[5] J. Hare, J. Davies, S. Samangooei, and P. Lewis, "Placing Photos with a Multimodal Probability Density Function", in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, 329.

[6] D. Jurgens, "That's What Friends are for: Inferring Location in Online Social Media Platforms Based on Social Relationships.", in *ICWSM*, 2013.

[7] E. Ilina, "A User Modeling Oriented Analysis of Cultural Backgrounds in Microblogging", *HUMAN JOURNAL*, 2012, 1 (4), 166–181.

[8] J. Mahmud, J. Nichols, and C. Drews, "Home Location Identification of Twitter Users", *ArXiv Preprint ArXiv:1403.2345*, 2014.

[9] V. Rakesh, C.K. Reddy, D.Singh, and M.S. Ramachandran, "Location-specific Tweet Detection and Topic Summarization in Twitter", in *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*. IEEE, 2013, 1441–1444.

[10] Y. Chen, J. Zhao, X.Hu, X. Zhang, Z. Li, and T.-S. Chua, "From Interest to Function: Location Estimation in Social Media.", in *AAAI*, 2013.

[11] J. Gelernter and W. Zhang, "A Learning Method to Geocode Location Expressions in Twitter Messages", *Journal of Spatial Information Science*, 2014.

[12] S. Hale, D. Gaffney, and M. Graham, "Where in the World are You? Geolocation and Language Identification in Twitter", *Professional Geographer*.

[13] J. Lingad, S. Karimi, and J. Yin, "Location Extraction from Disaster-related Microblogs", in *Proceedings of the 22nd International cCnference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 2013, 1017–1020.

[14] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge, "Supervised Text-based Geolocation Using Language Models on an Adaptive Grid", in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, 1500–1510.

[15] O. Van Laere, J. Quinn, S. Schockaert, and B. Dhoedt, "Spatially Aware Term Selection for Geotagging", *Knowledge and Data Engineering, IEEE Transactions on*, 2014, vol. 26, 1, 221–234.

[16] N. Ireson and F. Ciravegna, "Toponym Resolution in Social Media", in *The Semantic Web–ISWC 2010*, Springer, 2010, 370–385.

[17] R. Compton, M.S. Keegan, and J. Xu, "Inferring the Geographic Focus of Online Documents from Social Media Sharing Patterns", *ArXiv Preprint ArXiv:1406.2392*, 2014.

[18] Y. Takhteyev, A. Gruzd, and B. Wellman, "Geography of Twitter Networks", *Social networks*, 2012, vol. 34 (1), 73–81.

[19] J. Kulshrestha, F. Kooti, A. Nikravesh, and P.K. Gummadi, "Geographic Dissection of the Twitter Network.", in *ICWSM*, 2012.

[20] B.A. Olaniran, "Culture, Learning Styles, and Web 2.0", *Interactive Learning Environments*, 2009, vol. 17 (4), 261–271.

[21] B.E. Wiggins, *The Impact of Cultural Dimensions and the Coherence Principle of Multimedia Instruction on the Achievement of Educational Objectives within an Online Learning Environment*, PhD thesis, Indiana University of Pennsylvania, 2011.

[22] T.H. Silva, P. de Melo, J. Almeida, M. Musolesi, and A. Loureiro, "You are What you Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food & Drink Habits in Foursquare", *ArXiv Preprint ArXiv:1404.1009*, 2014.

[23] R. Garcia-Gavilanes, D. Quercia, and A. Jaimes, "Cultural Dimensions in Twitter: Time, Individualism and Power", *Proceedings of the 9 th AAAI ICWSM*, 2013.

[24] R. Garcia Gavilanes, Neil OHare, Luca Maria Aiello, and Alejandro Jaimes, "Follow my Friends this Friday! An Analysis of Human-generated Friendship Recommendations", in *Social Informatics*, 2013, Springer, 370–385.

[25] M. Giatsoglou, D. Chatzakou, and A. Vakali, "Community Detection in Social Media by Leveraging Interactions and Intensities", in *Web Information Systems Engineering–WISE 2013*, Springler, 2013, 57–72.

[26] Y. Artzi, P. Pantel, and M. Gamon, "Predicting Responses to Microblog Posts", in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2012, 602–606.

[27] J. Schantl, R. Kaiser, C. Wagner, and M. Strohmaier, "The Utility of Social and Topical Factors in Anticipating Repliers in Twitter Conversations", in *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 2013, 376–385.

[28] M.D.Myers and F.B. Tan, "Beyond Models of National Culture in Information Systems Research", *Advanced Topics in Global Information Management*, 2003, vol. 2, 14–29.

[29] R.D. Lewis, *When Cultures Collide: Managing Successfully Across Cultures*, Nicholas Brealey Publishing, 2000.

[30] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J.R. Curran, "Learning Multilingual Named Entity Recognition from Wikipedia", *Artificial Intelligence*, 2013, vol. 194, 151–175.

[31] D. Nadeau and S. Sekine, "A Survey of Named Entity Recognition and Classification", *Lingvisticae Investigationes*, 2007, vol. 30 (1), 3–26.

[32] T. Fawcett, "Roc Graphs: Notes and Practical Considerations for Researchers", *Machine learning*, 2004, vol. 31, 1–38.

[33] M. Sigala and O. Sakellaridis, "Web Users' Cultural Profiles and E-service Quality: Internationalization Implications for Tourism Web Sites", *Information Technology & Tourism*, 2004, vol. 7 (1), 13–22.